Программа учебной дисциплины
«Введение в Data Science»

Утверждена
Академическим советом ООП
.

Автор	Денике Екатерина Игоревна, преподаватель
Число кредитов	4
Контактная работа (час.)	16
Самостоятельная работа (час.)	152
Курс	Магистратура, 1 курс
Формат изучения дисциплины	С использованием онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения дисциплины «Извлечение и анализ интернет-данных» является формирование у студентов практических навыков анализа и извлечения данных и работы с ними.

В результате освоения дисциплины студент должен:
знать:
- основные методы анализа данных с помощью питон
- основные методы визуализации данных с помощью питон
- основные методы парсинга данных из интернета

уметь:
- извлекать информацию из различных интернет-ресурсов (как текст, так и изображения)
- обрабатывать информацию и данные, полученные из интернет-ресурсов
- визуализировать данные с помощью графических библиотек
- анализировать данные с помощью библиотек для анализа данных

Изучение данной дисциплины базируется на следующих дисциплинах:

- Линейная алгебра
- Математический анализ
- Основы программирования на Python

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- знать основы линейной алгебры и математического анализа;
- обладать навыками программирования в Python.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- Научно-исследовательский семинар

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Основы анализа данных в языке Python
Повторение основных функций и объектов языка Python. Обзор библиотек numpy, pandas на основе данных из соревнований платформы kaggle.com.

Тема 2. Визуализация данных в Python
Библиотеки matplotlib, seaborn, plotly. Продвинутые инструменты для анализа данных. Введение в визуальный анализ данных. Построение графиков, гистограмм, тепловых карт. Знакомство с порталом Открытых данных.

Тема 3. Парсинг открытых данных в различных форматах (xml/json/html)
Изучение языков и библиотек для работы с xml/json/html: lxml, XPath, XSLT, Beautiful Soup.

Тема 4. Основы машинного обучения и практика применения
Приведение текстовых данных к числовым с помощью OneHot- и TF-IDF кодирования, а также на основе представлений слов и текстов. Алгоритмы машинного обучения: линейная и логистическая регрессии, градиентный бустинг и нейронные сети.

Тема 5. Извлечение данных сайта Вконтакте и изучение влияния социальных сетей на поведение в реальной жизни
Изучаем возможности API сайта Вконтакте. Извлекаем информацию об интересах и демографии пользователей, на основании списка групп и поля “интересы”. Изучаем взаимосвязь интересов школьников с оценками.

III. ОЦЕНИВАНИЕ

Результирующая оценка по дисциплине выставляется из выполненных домашних работ и по результатам экзамена по формуле:

О_рез = 0,7 * О_накопл.+ 0,3 * О_экзамен

Накопленная оценка (О_накопл) рассчитывается как среднее значение оценок за все выданные домашние задания. Способ округления итоговой оценки по учебной дисциплине арифметический.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

Темы домашних заданий:

1. Основная функциональность Питона
2. Анализ и визуализация датафрейма
3. Извлечение данных из заданных ресурсов и их обработка

Оценочные средства для промежуточной аттестации:

Извлечение данных из популярных ресурсов и их последующий анализ

V. РЕСУРСЫ

5.1 Основная литература

1. Vanderplas JT. Python Data Science Handbook : Essential Tools for Working with Data [Internet]. Vol. First edition. Sebastopol, CA: O'Reilly Media; 2016. Available from: http://proxylibrary.hse.ru:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1425081&site=eds-live
2. Rossant, Cyrille. Learning IPython for Interactive Computing and Data Visualization, Packt Publishing Ltd, 2013. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1192643.

5.2 Дополнительная литература

1. Mark Lutz. Learning Python. 2008; Available from: http://proxylibrary.hse.ru:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsbas&AN=edsbas.6A062D5F&site=eds-live

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Anaconda Community	Свободно распространяемое лицензионное соглашение
2.	Python Software Foundation Python	Свободно распространяемое лицензионное соглашение
3.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	
4.	Microsoft Office Professional Plus 2010	

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.ru/

5.5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет;
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ноутбуками, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.